Physica A 449 (2016) 240–253

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Network structure exploration in networks with node attributes

Yi Chen [a], Xiaolong Wang [a,b], Junzhao Bu [a], Buzhou Tang [a,*], Xin Xiang [a]

[a] *Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China*

[b] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China*

HIGHLIGHTS

- Propose a novel Bayesian nonparametric (BNP) model.
- Explore network structural regularities.
- Handle networks with node attributes.
- Give stable experimental results.

ARTICLE INFO

Article history:
Received 11 February 2015
Received in revised form 29 November 2015
Available online 8 January 2016

Keywords:
Network structure
Structure exploration
Node attributes
Bayesian nonparametric model

ABSTRACT

Complex networks provide a powerful way to represent complex systems and have been widely studied during the past several years. One of the most important tasks of network analysis is to detect structures (also called structural regularities) embedded in networks by determining group number and group partition. Most of network structure exploration models only consider network links. However, in real world networks, nodes may have attributes that are useful for network structure exploration. In this paper, we propose a novel Bayesian nonparametric (BNP) model to explore structural regularities in networks with node attributes, called Bayesian nonparametric attribute (BNPA) model. This model does not only take full advantage of both links between nodes and node attributes for group partition via shared hidden variables, but also determine group number automatically via the Bayesian nonparametric theory. Experiments conducted on a number of real and synthetic networks show that our BNPA model is able to automatically explore structural regularities in networks with node attributes and is competitive with other state-of-the-art models.

1. Introduction

Complex networks provide a powerful way to represent various real world complex systems in nature and society, such as social systems [1,2], information systems [3,4], ecological systems [5], and others [6–8]. Exploring internal structures (also called structural regularities) is one of the most important tasks of complex network analysis. The task of network structure exploration is to determine how many groups in a complex network (group number) and how to partition objects

* Corresponding author.
E-mail addresses: chenyi@hitsz.edu.cn (Y. Chen), wangxl@insun.hit.edu.cn (X. Wang), bujunzhao@hitsz.edu.cn (J. Bu), tangbuzhou@gmail.com (B. Tang), xiangxin@hitsz.edu.cn (X. Xiang).

http://dx.doi.org/10.1016/j.physa.2015.12.133

(i.e., nodes) into different groups (group partition). Fundamentally, nodes within a group have common characteristics. For example, in social networks, friends who attended the same school may form a group. There are many types of structures including assortative structure (i.e., the community structure), disassortative structure (e.g., the bipartite structure) [9], mixture structure [10] and so on. The assortative structure is a type of network structure in which most edges are within a group, while the disassortative structure is a type of network structure in which most edges are across groups. The mixture structure is a type of network structure that contains both assortative and disassortative structures. A large number of effective models have been proposed to explore these types of structures during the last several years, and a detailed survey about them was presented by Fortunato [11]. Most of them are only based on links between nodes.

However, besides links, nodes in real world networks may have many attributes. For example, users in an online social network usually have a profile to store personal information like age, gender and ethnic background and so on. The attributes of nodes may also provide useful information for structure exploration. For example, users with same attributes in an online social network are potential to group together. How to use links between nodes and node attributes to explore structural regularities simultaneously? Recently, a few models have been proposed for network structure exploration in networks with node attributes. They may fall into two categories: heuristic measure-based methods [12–14] and probabilistic inference-based models [10,15–21]. The heuristic measure-based models unify links between nodes and node attributes via a heuristic function. Zhou et al. [13] proposed a neighborhood random walk distance to measure both link similarities and attribute similarities to cluster a graph with attributes. Ruan et al. [12] used content information to determine the strength of links between nodes. Akoglu et al. [14] introduced a parameter-free identification of cohesive subgroups (PICS) in attributed graphs via a total encoding cost function of both model description cost and data description cost. The probabilistic inference-based models use a probabilistic function to describe the generating process of networks with node attributes and explore their network structure based on the model parameters. Cohn and Hofmann [19] first proposed probabilistic model based on Probabilistic HITS (PHITS) and Probabilistic Latent Semantic Analysis (PLSA) to model a document collection with inter-connectivity and contents, in which PHITS generates the inter-connectivity and PLSA generates the document contents. By replacing PLSA by latent Dirichlet allocation (LDA) and PHITS by LDA-link, Erosheva et al. [20] proposed a Bayesian model to generate a scientific citation network. Yang et al. [15,17] developed a generative link and discriminative content (DC) model to combine link and content for community detection. Chai et al. [10] first proposed a popularity–productivity stochastic block (PPSB) model for general structure detection, and then extended it to identify structural regularities in text-associated networks. The main disadvantage of these generative probabilistic models lies in that their group number should be assigned at first. To determine the group number in networks with node attributes automatically, Yang et al. [16] adopted cross-validation. Recently, The Bayesian nonparametric (BNP) theory [22], designed to automatically determine the component number of generative probabilistic models, has been introduced to determine the group number in networks with node attributes. For example, Sinkkonen et al. [21] proposed a simple BNP model for enriched graphs. Duan et al. [18] proposed a BNP model to simultaneously detect communities and topics in text-augmented social networks, in which the Dirichlet Process Mixture model and the Hierarchical Dirichlet Process model are used to automatically determine the numbers of communities and topics, respectively.

In this paper, we propose a novel BNP model based on Newman's mixture model and the Bayesian nonparametric theory to explore structural regularities in networks with node attributes, called Bayesian nonparametric attribute (BNPA) model. Firstly, Newman's mixture model (NMM) [9], a model for structural regularity exploration in networks without node attributes, is extended for networks with node attributes, called Newman's mixture model with attributes (NMMA). Then NMMA is further extended using the Bayesian nonparametric theory. The BNPA model does not only take full advantage of both links between nodes and node attributes for group partition via shared hidden variables, but also determine group number automatically via the BNP theory. Experiments conducted on a number of synthetic and real datasets show that the BNPA model is able to automatically explore structural regularities and achieves better performance than other state-of-the-art models on most of the networks.

The remainder of this paper is organized as follows. Section 2 presents the BNPA model. Experiments are presented in Section 3. Section 4 draws conclusions.

2. Bayesian nonparametric attribute (BNPA) model

An undirected network with N nodes is represented by an $N \times N$ adjacency matrix A, where $A_{ij} = 1$, $(1 \le i, j \le N)$ if there is a link from node i to node j and 0 otherwise. The node attributes are represented by a $N \times M$ matrix W, where M is the dimension of attributes, $W_{it} = 1$, $(1 \le i \le N, 1 \le t \le M)$ if node i has attribute t and 0 otherwise. For convenience, we use $N(i)$ to denote the links of node i and use $M(i)$ to denote the attributes of node i. In this section, we first introduce NMMA, then BNPA.

2.1. Newman's mixture model for networks with attributes (NMMA)

The NMM model is a classic generative probabilistic model for network structure exploration in networks with only links. A network with K (a predefined number) groups can be generated by the NMM model with two parameters π_k and θ_{kj}, where π_k denotes the probability of a node in group $k (k \in \{1, 2, \dots, K\})$ and θ_{kj} denotes the probability of a link from a